

Introduction to Logistic Regression in R

(with case studies on the phonological organization of mental lexicon)

T. Florian Jaeger and Peter Graff

1: University of Rochester

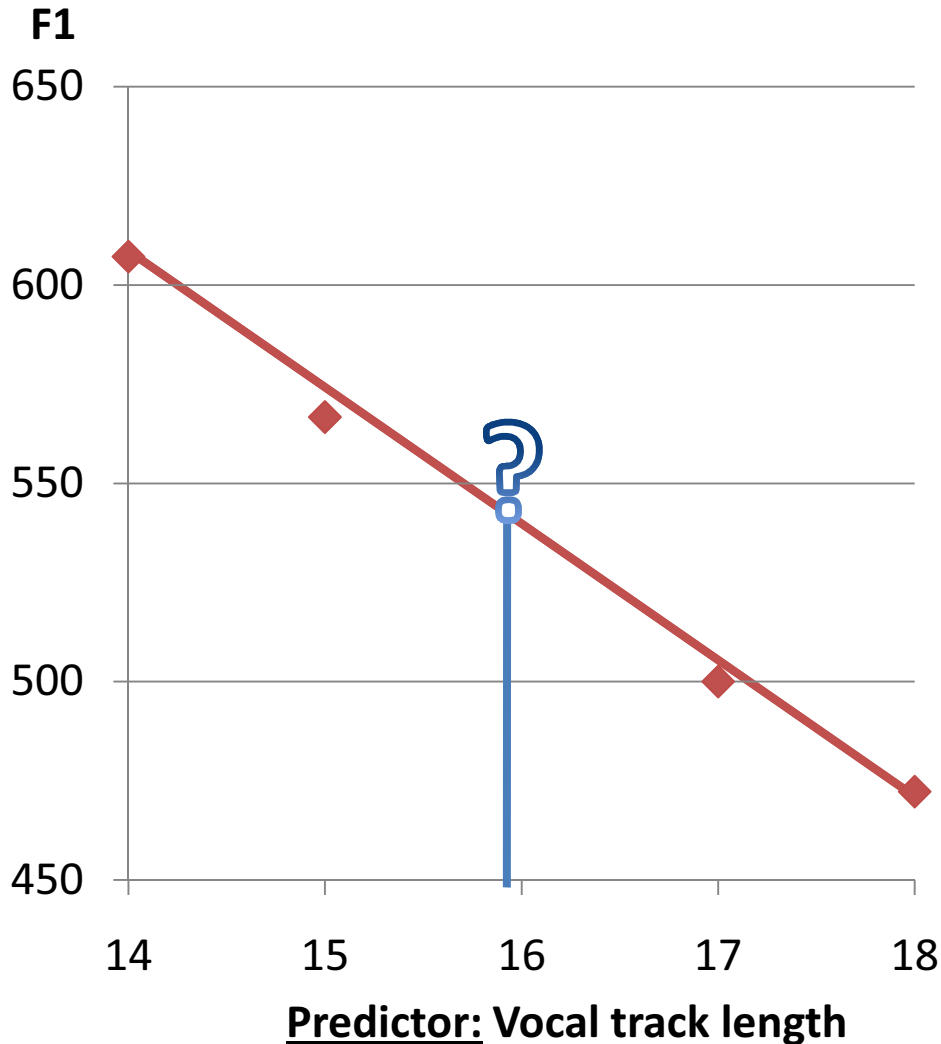
2: Massachusetts Institute of Technology

Intro

- **Part I:** Geometrical view of things
- **Part II:** quick intro to GLM and GLMM
- **Part III:** case study using logistic regression to study similarity avoidance in the mental (phonological) lexicon
- **Part V:** Discussion – **but please feel free to ask questions any time**

Predicting unobserved data points

Outcome:



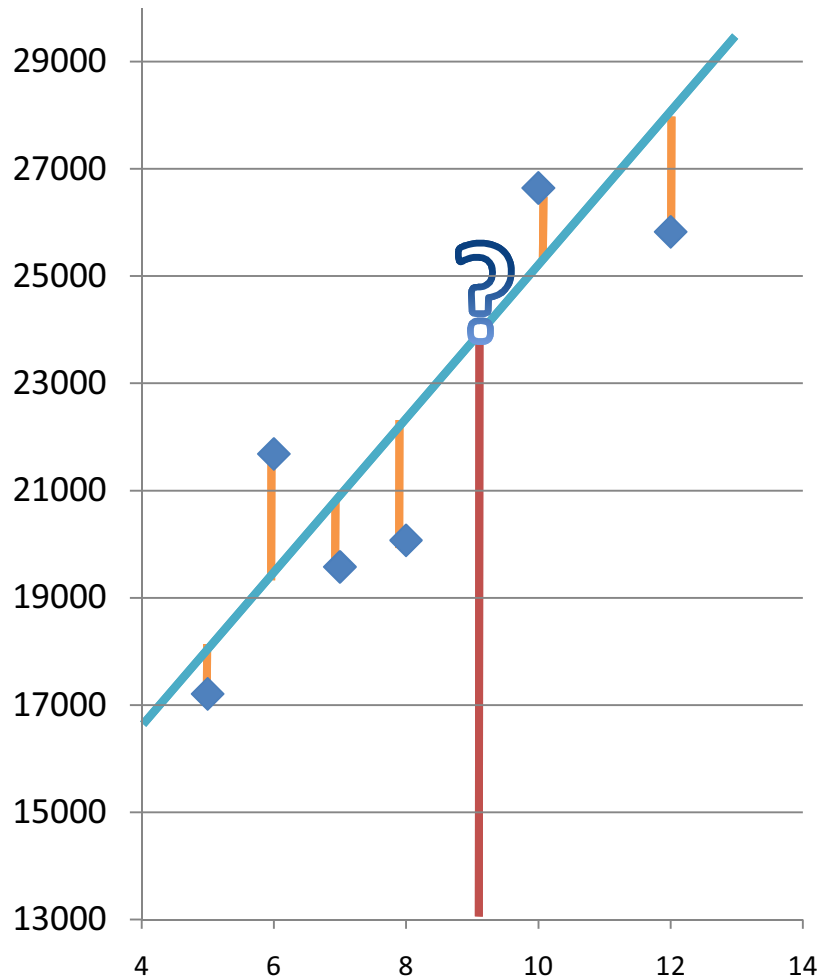
- Assume we want to predict **F1** from **vocal tract length** based on a limited sample.
- Fitting a **linear model**:
 $F1 = \text{intercept} + \beta \cdot \text{VocTL}$
- NB: what does linear mean here?

$$\begin{aligned} F1 = & \text{intercept} \\ & + \beta_1 \cdot \text{VocTL} \\ & + \beta_2 \cdot \text{VocTL}^2 \end{aligned}$$

Predicting vs. Evaluating Significance

- You can think of regression in at least two ways:
 - Building a predictive model (based on observed data, you want to be able to make best guesses about future observations)
 - Testing whether a predictor affects an outcome (significantly)
- These views are related: in both cases, we need to find the best β

What constitutes the best guess?



- Orange lines = Error in prediction
- This is minimizing the squared deviations from the line (squared error)
- Do you notice something? How does regression differ from correlation?

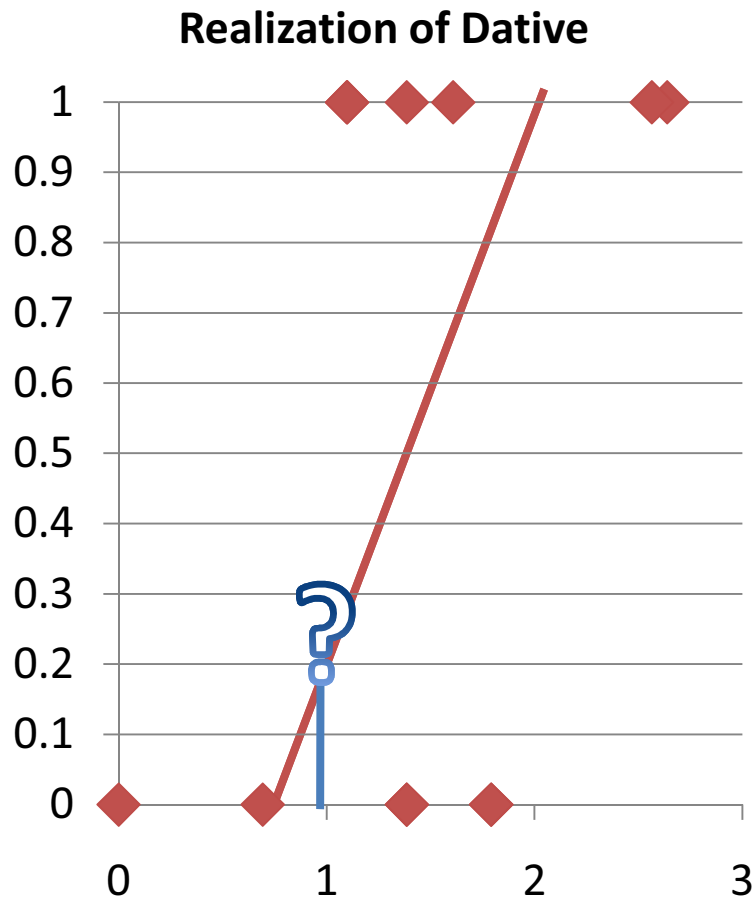
Limitations of Linear Model

- Assumptions:
 - Linearity in coefficients
 - normally distributed outcome (or error around outcome) → non-continuous outcomes are usually **not** normally distributed
- But many/most of the outcomes of interest to linguists are categorical outcomes

Categorical outcomes

- Choices in syntactic/morpho-syntactic variation:
 - Dative alternation, Heavy NP shift, Particle shift
 - *that*-omission, optional case-marker omission, optional, clitic doubling, argument drop, ellipsis,
 - Auxiliary contraction, phone deletion
- Forced-choice experiments: Grammaticality; yes/no question; multiple choice questions
- Eye-tracking: fixations on one of several referents
- Typological work: absence/presence of grammatical features, words, etc. across languages

Can a linear model do the job?



- Predicting Realization of Dative (NPNP vs. NPPP) from Length of Theme (log).
 - Predicts impossible values (<0 , >1)
 - Make unlikely assumptions about distribution of variable

Let's take a step back

- What does a linear model actually predict?
 - The linear predictor (the formula) predicts the *mean* of the outcome ...
 - ... and then we add some noise
- What do we want to predict for binary outcomes?
 - The probability of outcome A over outcome B

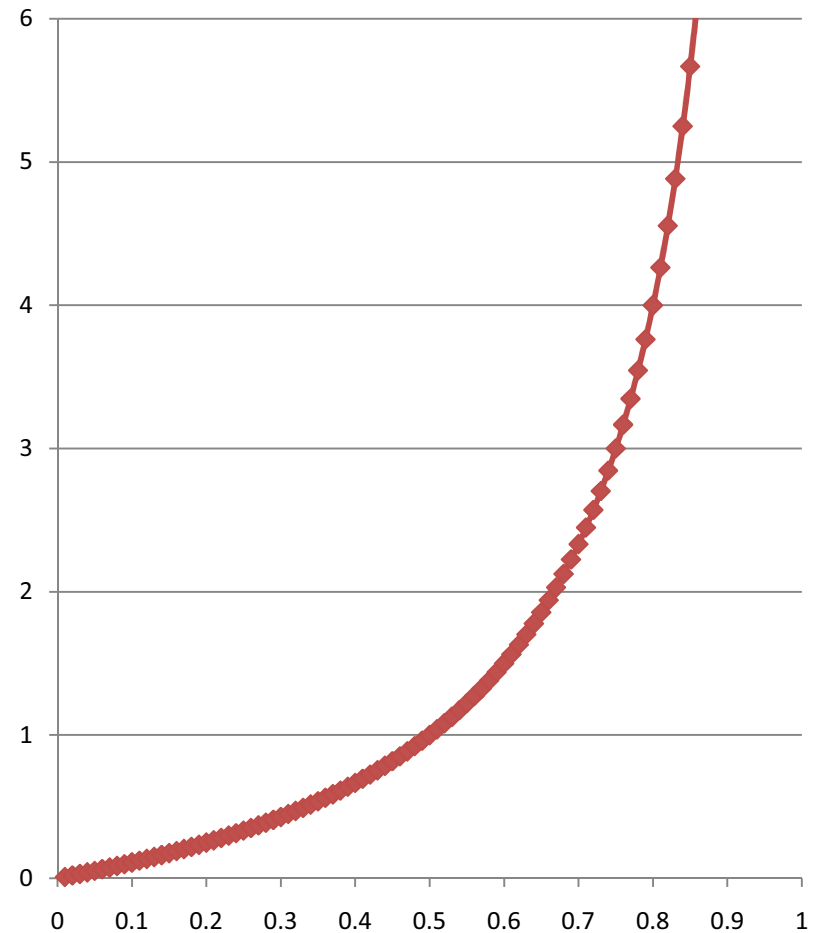
Odds

- Probabilities range between 1 and 0
- We can transform them into a measure ranging from 0 to infinity

- Odds

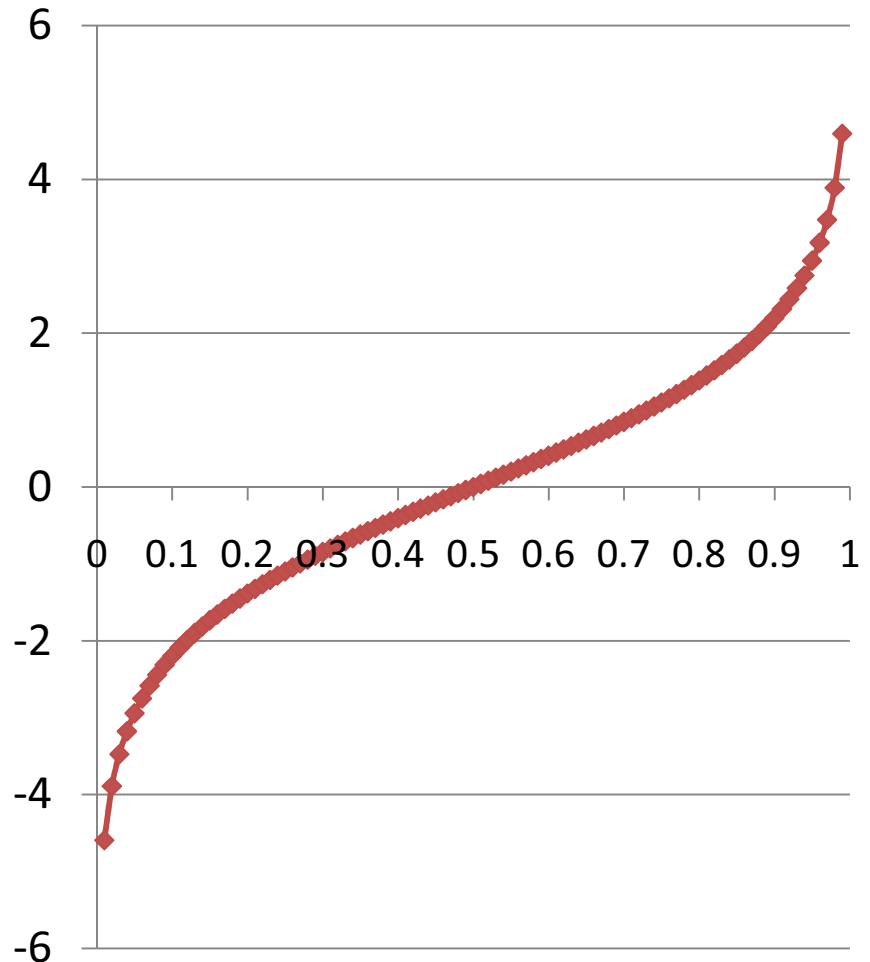
$$o = (p/1-p)$$

- $p < .5$, $0 < o < 1$
- $p = .5$, $o = 1$
- $p > .5$, $o > 1$



Log Odds

- We take the natural logarithm of the odds ratio (a.k.a. **logit**)
 - A value of 0 at $p=.5$
 - Probabilities with the same distance above and below .5 have the same logits but different signs.



Case Study: The OCP in Javanese

- Prohibitions against or under-attestation of combinations of similar sounds are commonly known as OCP effects
- Javanese is known for co-occurrence restrictions on similar sounds within words
(Mester, 1986)
- We'll cover important aspects of logistic regression analysis while constructing a model of the OCP in Javanese

Generative Potential

- A priori we might expect that languages make use of their full generative potential.
- If a language allows words of the general shape CVCVC, every possible permutation of consonants in the C-slots could be a word.
- Nonetheless, the majority of possible permutations of consonants is unattested.

Javanese

- 30% of all Javanese roots are CVCVC
- Of 9,261 (21^3) theoretically possible CVCVC roots 1,913 (20%) are attested
(Uhlenbeck, 1978)
- We generated every possible consonant triplet of Javanese and annotated it 1 if attested and 0 if unattested.

Javanese Corpus Input File

template	attestation	type frequency
tSVIVb	0	0
tSVIVtS	0	0
tSVIVd	0	0
tSVIVg	0	0
tSVIVh	1	5
tSVIVj	0	0
tSVIVk	1	4
tSVIVl	1	1
tSVIVm	1	2

Javanese Corpus Input File

template	attestation	type frequency
tSVIVb	0	0
tSVIVtS	0	0
tSVIVd	0	0
tSVIVg	0	0
tSVIVh	1	5
tSVIVj	0	0
tSVIVk	1	4
tSVIVl	1	1
tSVIVm	1	2

Javanese C_1VC_2 co-occurrence restrictions (Mester, 1986)

$C_2 \backslash C_1$	p	b	m	w	t	d	T	D	r	l	n	ɲ	s	c	j	y	k	g	ɔ	h	
p		X	X	X																	
b	X			X	X																
m	X	X																		X	
w	X	X	X					X						X							
t						X	X						X	X	X						
d					X			X	X							X					
T																					
D					X	X	X		X	X						X					
r								X	X	X	X										
l										X											
n																					
ɲ														X	X	X					
s													X	X		X	X				
c				X	X																
j													X	X	X						
y																					
k																				X	X
g																				X	X
ɔ																					

Similarity is dispreferred

Javanese C_1VC_2 co-occurrence restrictions (Mester, 1986)

$C_1 \backslash C_2$	p	b	m	w	t	d	T	D	r	l	n	ɲ	s	c	j	y	k	g	ɔ	h
p		X	X	X																
b	X		X	X																
m	X	X																X		
w	X	X	X				X							X						
t					X	X					X	X	X							
d					X		X	X					X	X						
T																				
D					X	X	X		X	X						X				
r							X	X		X	X									
l																				
n																				
ɲ													X	X	X					
s					X	X							X	X	X	X				
c													X	X	X	X				
j													X	X	X	X				
y																				
k																		X	X	
g																		X	X	
ɔ																			X	X

Identity is favored

What we're doing vs. O/E's

- O/E's have become a standard measure of sound co-occurrence in phonology
- The problem with O/E's is...
 - Inflation of Type I Error
 - Difficult to get E's right if several variables play a role (e.g. identity)
 - Difficult to tease apart the relative contributions of the variables influencing the E's

To ask if the OCP has shaped
the lexicon of Javanese...

- Phonotactics and a preference for identity might obscure the result!
- We need to control for ***both*** identity ***and*** occurrence effects

Control 1: Occurrence Restrictions

- Restrictions on the occurrence of certain sounds in certain positions affect the probability of a form's attestation.
 - Javanese tVbVw is unattested. It is possible to attribute the non-attestation to co-occurrence restrictions on labials but there is no word that ends in /w/.
- Occurrence as positional frequency.

Positional Frequency Factors (3)

- One frequency factor per consonant-slot.
= number of attested templates in which i^{th} consonant of the template occurs in position i .
- **Example** (235 attested templates start with /t/)

tVtVk

Frequency. C_1 in C_1 = 235

Control 2: Identity

- Several languages with strong OCP effects allow for total identity between consonants.

(McEachern 1997, Gallagher and Coon 2009)

- Identity might even be preferred cross-linguistically.

(Zuraw, 2002)

Identity Factors (3)

- One factor per pair of C-slots
= 1 if C_i and C_j are identical
= 0 otherwise

- *Example*

tVtVk

Identity.C₁inC₂ = 1

Outline

- ***Nested and Non-nested Model Comparison***
 - Are there OCP effects after other factors have been controlled for?
 - Does the OCP have to refer to individual features?
 - Does the OCP require a notion of locality?
 - How does our model compare to other models of the OCP?
- ***Effect size***
 - Does the strength of OCP effects differ for different phonological features?

Question 1

Are there OCP effects after other factors have been controlled for?

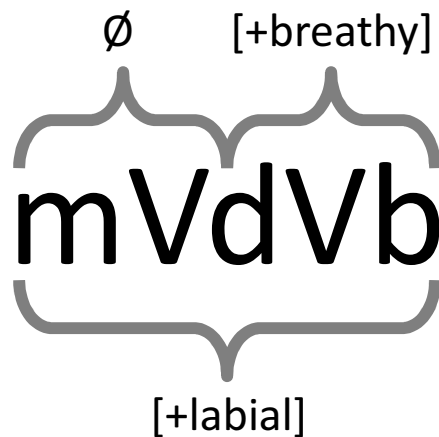
Feature System

- Place
 - labial
 - alveolar
 - post-alveolar
 - palatal
 - velar
 - glottal
 - retroflex
- Manner+
 - lateral
 - rhotic
 - nasal
 - strident
 - continuant
 - sonorant
 - approximant
- Laryngeal
 - breathy

OCP Model 1: Sum of Matches

- One variable
= Sum of feature matches between C_1 & C_2 , C_2 & C_3
and C_1 & C_3

- **Example**



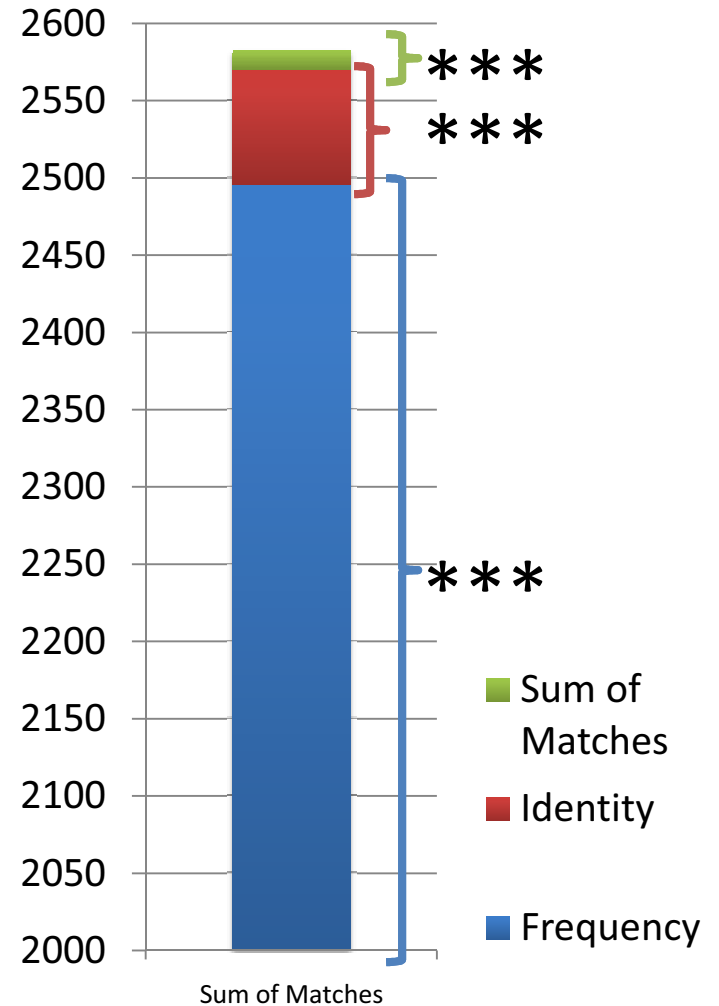
Sum.of.Matches = 2

Fitting the Model

```
lrm (attestation~  
  OCP { sum.of.matches+  
  Frequency { C1inC1.fq+C2inC2.fq+  
  Identity { C3inC3.fq+  
  identity.C1C2+  
  identity.C1C3+  
  identity.C2C3  
  , data=jav) ->jav1
```

anova (jav1) gives us...

Factor	Chi-Square	d.f.	P
Sum.of.Matches	13.21	1	0.0003
identity.C1C2	45.92	1	<.0001
identity.C1C3	21.08	1	<.0001
identity.C2C3	7.49	1	0.0062
C1inC1.fq	903.49	1	<.0001
C2inC2.fq	241.88	1	<.0001
C3inC3.fq	1350.04	1	<.0001
TOTAL	1802.95	7	<.0001




Question 2

Does the OCP have to refer to individual features?

OCP Model 2: Feature Specific OCP

- One variable per feature
= 1 if one pair of consonants match for that feature
= 2 if all three consonants match for that feature
- *Example*

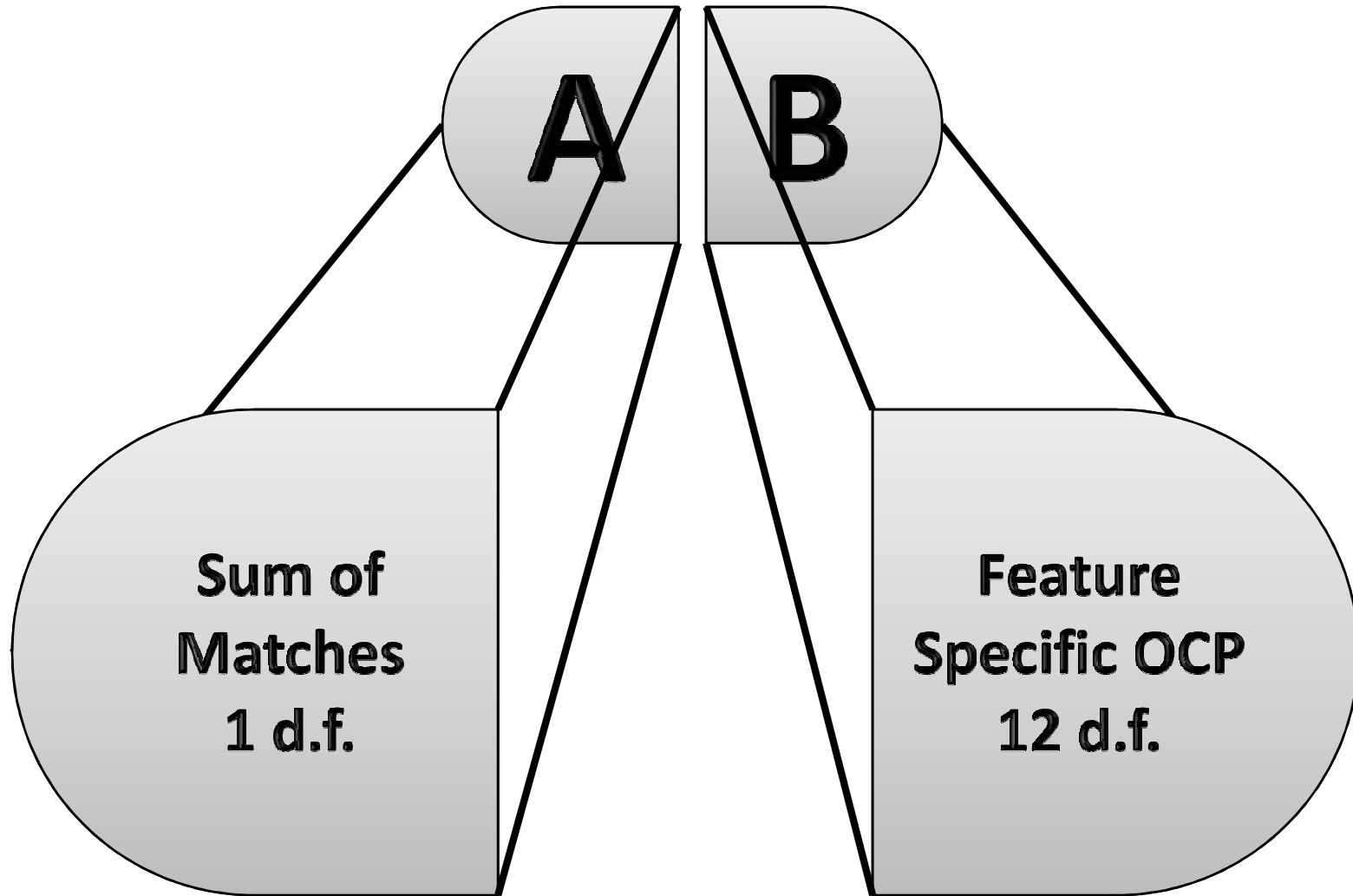
mVdVt



[+alveolar]

OCP.alveolar = 1

Non-nested Model Comparison



Fitting the Model

```
lrm(attestation~  
OCP.labial+OCP.alveolar+  
OCP.retroflex+  
OCP.post.alveolar+  
OCP.palatal+OCP.velar+  
OCP.glottal+OCP.nasal+  
OCP.rhotic+OCP.strident+  
OCP.lateral+OCP.breathy+  
Sum.of.Matches+  
C1inC1.fq+C2inC2.fq+C3inC3.fq+  
identity.C1C2+identity.C1C3+  
identity.C2C3  
, data=jav) -> jav2
```

OCP (new) {

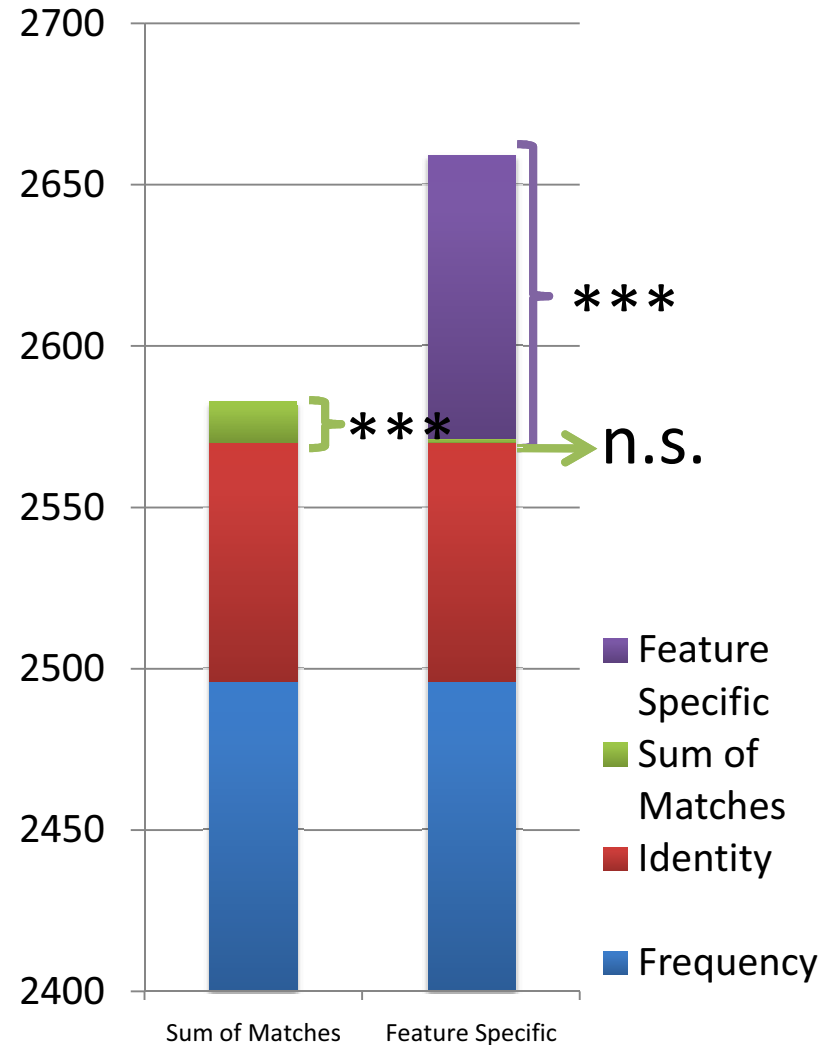
OCP (old) {

Frequency {

Identity {

anova (jav2) gives us...

Factor	Chi-Square	d. f.	P
OCP.labial	23.68	1	<.0001
OCP.alveolar	6.09	1	0.0136
OCP.retroflex	0.17	1	0.6790
OCP.post.alveolar	2.47	1	0.1161
OCP.palatal	1.19	1	0.2759
OCP.velar	3.63	1	0.0566
OCP.glottal	19.72	1	<.0001
OCP.nasal	3.96	1	0.0467
OCP.rhotic	9.78	1	0.0018
OCP.lateral	12.59	1	0.0004
OCP.breathy	0.98	1	0.3225
OCP.strident	3.97	1	0.0464
Sum.of.Matches	1.25	1	0.2632
identity.C1C2	60.3	1	<.0001
identity.C1C3	6.52	1	0.0107
identity.C2C3	0.28	1	0.594
C1inC1.fq	873.22	1	<.0001
C2inC2.fq	225.13	1	<.0001
C3inC3.fq	1253.92	1	<.0001
TOTAL	1812.77	19	<.0001



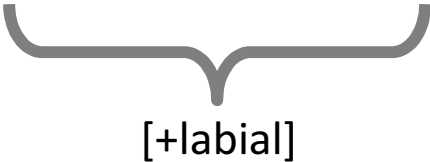
Question 3

Does the OCP require a notion of locality?

OCP Model 3: Non-local OCP

- One *additional* variable per feature
= 1 if C1 and C3 match for that feature
- *Example*

mVdVb



[+labial]

Non.Local.OCP.labial = 1

Fitting the Model

```
lrm(attestation~  
OCP.labial+OCP.alveolar+ OCP.retroflex+  
OCP.post.alveolar+ OCP.palatal+OCP.velar+  
OCP.glottal+OCP.nasal+  
OCP.rhotic+OCP.strident+  
OCP.lateral+OCP.breathy+  
OCP.labial+OCP.alveolar+ OCP.retroflex+  
OCP.post.alveolar+ OCP.palatal+OCP.velar+  
OCP.glottal+OCP.nasal+  
OCP.rhotic+OCP.strident+  
OCP.lateral+OCP.breathy+  
C1inC1.fq+C2inC2.fq+C3inC3.fq+  
identity.C1C2+identity.C1C3+  
identity.C2C3  
, data=jav) -> jav3
```

OCP (old)

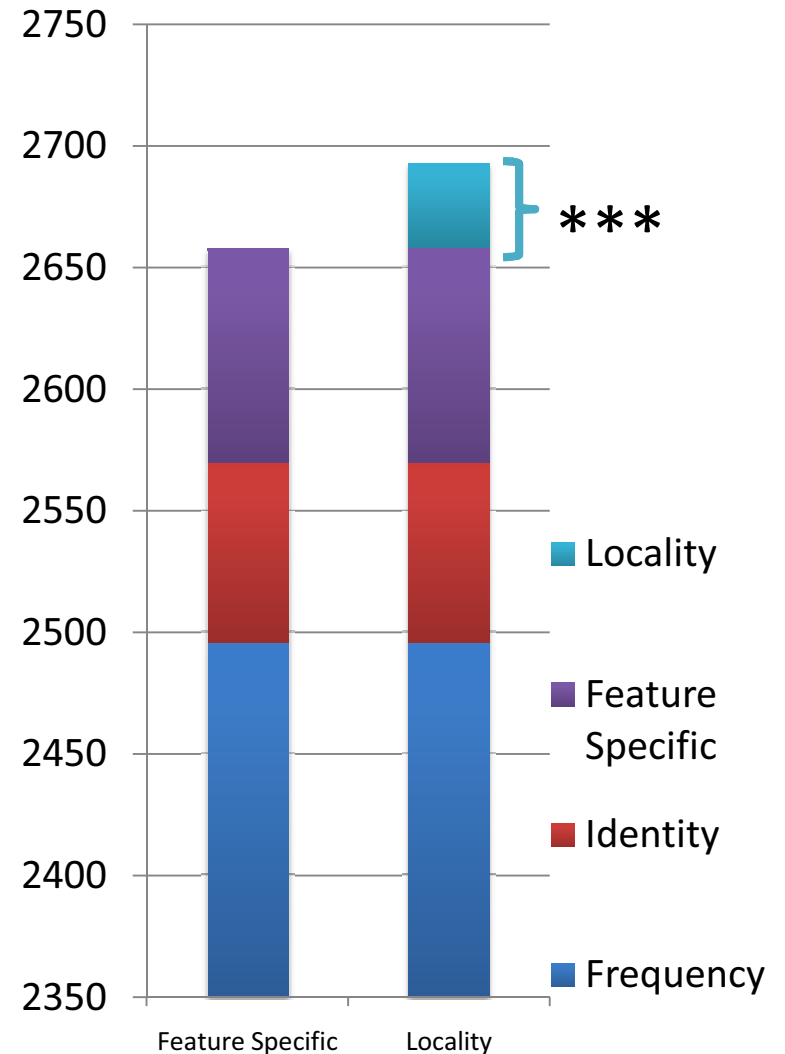
OCP (new)

Frequency

Identity

anova (jav3) gives us...

Factor	Chi-Square	d.f.	P
Non.Local.OCP.labial	3.97	1	0.0463
Non.Local.OCP.alveolar	5.3	1	0.0213
Non.Local.OCP.retroflex	0.01	1	0.9025
Non.Local.OCP.post.alv	0.02	1	0.8889
Non.Local.OCP.palatal	0.67	1	0.4124
Non.Local.OCP.velar	5.78	1	0.0162
Non.Local.OCP.glottal	0.02	1	0.8851
Non.Local.OCP.nasal	0.75	1	0.3858
Non.Local.OCP.rhotic	17.76	1	<.0001
Non.Local.OCP.lateral	0.7	1	0.4028
Non.Local.OCP.breathy	0.14	1	0.7110
Non.Local.OCP.strident	0.03	1	0.8704
OCP	94.52	12	<.0001
identity.C1C2	57.4	1	<.0001
identity.C1C3	10.9	1	0.0010
identity.C2C3	0	1	0.9800
C1inC1.fq	861.73	1	<.0001
C2inC2.fq	227.67	1	<.0001
C3inC3.fq	1197.4	1	<.0001
TOTAL	1727.88	30	<.0001



Conclusion

- The lexicon is shaped by the OCP even after controlling for language-specific phonology in a systematic way.
- We have shown evidence that OCP effects are Feature-Specific and require a notion of Locality